ABSTRACT

Artificial Intelligence is now a days gaining immense importance and is becoming a key technology in many fields ranging from banking industry, to travel industry, to communication industry, and to robotic industry. The use of Artificial Intelligence in medical diagnosis too is becoming increasingly common and has been used widely in the diagnosis of cancers, tumors, hepatitis, lung diseases, etc... The main aim of this paper is to build an Artificial Intelligent System that after analysis of certain parameters can predict that whether a person is diabetic or not. Diabetes is inability of body to manage the levels of sugar in the blood. It being one of the most chronic diseases around the world causes around 3.8 million deaths every year. Authors have identified 10 parameters that play an important role in diabetes and prepared a rich database of training data which served as the backbone of the prediction algorithm. Keeping in view this training data authors developed a system that uses the naïve-Bayes classification algorithm to serve the purpose. When the parameters of the test data are fed to the system, it anticipates & classifies the test data into one of the two categories viz diabetic & not diabetic. The performance of AI method when compared with the medical diagnosis system was found to be 95%. This system can be used to assist medical programs especially in geographically remote areas where expert human diagnosis not possible with an advantage of minimal expenses and faster results.

Keywords: Artificial Intelligence, Data Mining, Machine Learning, Diabetes, Naïve Bayes classifier, Medical Diagnosis.

1. INTRODUCTION

Diabetes is a chronic condition that occurs when the body cannot produce enough or cannot effectively use insulin [1]. Insulin is a hormone produced by pancreas and is needed by our body to metabolize glucose. When the glucose level in body is not metabolized properly it keeps on ci...
Now the "naïve" conditional independence assumptions come into play: assume that each feature is conditionally independent of every other feature \( F_i \) for \( i \) not equal to \( j \). This means that:

\[
p(F_i | C, F_j) = p(F_i | C)
\]

for \( i \) not equal to \( j \) and so the joint model can be expressed as

\[
p(C, F_1, F_2, ..., F_n) = p(C) p(F_1 | C) p(F_2 | C) p(F_3 | C) ....
\]

\[
= p(C) \prod p(F_i | C)
\]

This means that under the above independence assumption the conditional distribution over the class variable \( C \) can be expressed like this:

\[
p(C, F_1, F_2, ..., F_n) = 1/Z p(C) \prod p(F_i | C)
\]

where \( Z \) (the evidence) is a scaling factor dependent only on \( F_1,F_2,F_3,.....F_n \) i.e. a constant if the value of the feature variable \( s \) are known.

3. PREVIOUS WORK

It has been noted that the Machine learning algorithms are increasingly being used in solving problems in Medical Domains such as in Oncology[5,6,7,8], Urology[8,9], Hepatitis[9,13], Liver Pathology[10], Cardiology[15,16,17], Gynecology[18], Thyroid disorders[11,12], Tuberculosis[22], Neuropsychology[19], Perinatology [20] etc. Various algorithms have been used in different domains however the naïve Bayes algorithms have been noted to outperform most of the advanced and sophisticated algorithms in both medical diagnostic problems and in problems of non-medical domain [14]. Kononenko and others did a comprehensive comparison of naïve Bayes algorithm with six other algorithms and found that the naïve Bayes algorithm outperformed all other algorithms in 5 out of 8 problems of medical diagnostic domain. In a study, the Inductive Logic programming algorithms achieved a minimal classification accuracy of 12% to 29%, while the naïve Bayes algorithm for similar problem achieved an accuracy of 35% [14]. Yet in another comparison between naïve Bayes and modern decision tree algorithms like C4.5 (Quinlan 1993) has proved that the naïve Bayes prediction capabilities are equally good as C4.5 (Langley, Iba, & Thomas 1992; Pazzani 1996; Kononenko 1990) [21].

4. PARAMETERS USED IN ESTIMATION

Since India is having the highest Diabetic population in the world so it was easy to collect the data about the patients who suffered from this disease. After a detailed study, authors identified ten best physiological parameters for the study which were so chosen that the values for them could be easily determined and could be assigned discrete values, for the sake of maintaining consistency. Table-I summaries the parameters chosen and their allowed values. A dataset of 415 cases was prepared by collecting the data randomly from different sections of the society with an aim to have a variety in the dataset. To maintain accuracy and to avoid errors, considerable care was taken to ensure that the database had correct values.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Allowed Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Age of the subject</td>
<td>Discrete Integer values</td>
</tr>
<tr>
<td>Family History</td>
<td>Whether any family member of the subject is suffering/ was suffering from diabetes</td>
<td>Yes or No</td>
</tr>
<tr>
<td>Sex</td>
<td>Whether male or female</td>
<td>Male or Female</td>
</tr>
<tr>
<td>Smoking</td>
<td>Whether the subject does smoking or not</td>
<td>Yes or No</td>
</tr>
<tr>
<td>Drinking</td>
<td>Whether the subject does drinking or not</td>
<td>Yes or No</td>
</tr>
<tr>
<td>Fatigue</td>
<td>Does a person feel tired after doing a little work?</td>
<td>Yes or No</td>
</tr>
<tr>
<td>Thirst</td>
<td>Whether the subject frequently feels a strong desire to drink water. i.e how many times the subject drinks water.</td>
<td>Discrete Integer values</td>
</tr>
<tr>
<td>Frequency of Urination</td>
<td>How many times the subject passes urine in a day</td>
<td>Discrete Integer values</td>
</tr>
<tr>
<td>Height</td>
<td>Height of the subject</td>
<td>Discrete floating point values</td>
</tr>
<tr>
<td>Weight</td>
<td>Weight of the subject</td>
<td>Discrete floating point values</td>
</tr>
</tbody>
</table>

5. IMPLEMENTATION

As per the Conditional independence assumption of Bayes theorem, the presence or absence of some parameters of a class is independent to the presence or absence of some other parameters, making each parameter’s contribution independent to the final result. The authors calculated the individual probability of all the variables for both Diabetic=’Yes’ & Diabetic = ‘No’. For instance for a parameter “Frequency of Urination”, the probability of both Diabetic =’Yes’ & Diabetic = ‘No’ is calculated as:

\[
P(\text{Diabetic}=’Yes’) \text{ given } \text{’Frequency of Urination’} = \text{’Value from Test Data’} \text{ and } P(\text{Diabetic}=’No’) \text{ given } \text{’Frequency of Urination’} = \text{’Value from Test Data’}.\]

In the similar way, the probabilities of all the parameters and stored there individual contribution to the final result in different variables can be calculated. To deal with the condition of zero probability values for some parameter, the authors made use of Laplace Correction. At last the consolidation to the contribution of all the individual variables, according to the test data gets classified into one of the two categories viz Diabetic or Not Diabetic. The development of the system is done using
MATLAB with SQL server as database. The experimental set up in execution is shown in the Figure-1.

**Fig. 1: MATLAB Program in execution**

6. CONCLUSION

This naïve Bayes classifier based system is very useful for diagnosis of diabetes. The reliability of the system was evaluated by computing the mean absolute error between the predicted values and exact values the cases. The results suggest that this system can perform good prediction with least error and finally this technique could be an important tool for supplementing the medical doctors in performing expert diagnosis. In this method the efficiency of Forecasting was found to be around 95%. Its performance can be further improved by identifying & incorporating various other parameters and increasing the size of training data.

7. ACKNOWLEDGEMENT

This paper has benefited by the discussions with people in the Area of Artificial Intelligence, Medical Sciences and Academicians up to a greater extent

8. REFERENCES

[3] Optimal naïve Bayes harry zahang